White Paper Report

Report ID: 103954

Application Number: HW5001911

Project Director: Julia Flanders (j.flanders@neu.edu)

Institution: Northeastern University

Reporting Period: 9/1/2011-9/30/2013

Report Due: 12/31/2013

Date Submitted: 4/28/2014

Knowledge Organization and Data Modeling in the Humanities

Fotis Jannidis, University of Würzburg

Julia Flanders, Northeastern University

On March 14-16, 2012, the Women Writers Project at the Brown University and the Center for Digital Editions at the University of Würzburg organized a three-day workshop with generous sponsorship from the National Endowment for the Humanities and the Deutsche Forschungsgemeinschaft. The event brought together a group of about 35 on-site participants and a community of about 85 virtual participants who followed the live stream and the Twitter feed. Some of these virtual participants asked questions or contributed comments, and some of them helped communicate the substance of the event to a wider twitter audience.

The design of the event was aimed at creating the conditions for a substantive, wide-ranging, interdisciplinary, expert discussion of humanities data modeling grounded in many relevant fields. To accomplish this, we invited participants to make several different kinds of contributions, including papers on theoretical approaches, case studies grounded in specific projects, focused panel discussions on theory and on pedagogy, and also more general discussion at intervals throughout the event. You can find slides, papers, and videos of the event online.

The following white paper tries to sum up important topics and problems which came up in the presentations and discussions and to outline some general aspects of data modeling in digital humanities. Its aim is to offer a reference point for further research and to stimulate discussion on a topic crucial to Digital Humanities.

Introduction

The question of what would constitute "theory" for digital humanities is contested and somewhat fraught. From the perspective of humanists for whom "theory" means the cultural and critical theory that has been naturalized in humanities departments during the past thirty years, Alan Liu's recent question, "where is cultural criticism in the digital humanities?" reflects a general sense that the digital humanities has done too little to theorize its work.[1] For others working in the field, the place to look for digital humanities "theory" would be rather in the philosophical, logical, computational, and mathematical systems that undergird the representational structures we use. Thus the statement that "the database is the theory" could be unpacked to refer to work that examines theories of database design, first-order logic, [etc. e.g. from Renear, etc.].[2] Both of these positions create an artificial distance: the assumption that

[1] Benjamin M. Schmidt: Theory First. In: Journal of Digital Humanities 1,1 (2011). http://journalofdigitalhumanities.org/1-1/theory-first-by-ben-schmidt/

[2] Jean Bauer: Who you calling untheoretical? In: Journal of Digital Humanities 1,1 (2011).

digital humanities theory will be very close to established cultural theories ignores the fact that digital humanities is bringing together two very different fields both with their own and very different theoretical traditions and that both traditions will have to be considered and, if possible, merged in this new field.

The claim that the database is the theory on the other hand is in conflict with the meaning of 'theory'. Theory is usually the theory of something, trying to spell out the basic concepts relevant in the praxis of doing something. An interpretation and a theory of interpretation for example are therefore very different even if the interpretation is very theory-informed - they try to achieve different goals and probably they use different means to do so. Therefore a theory of digital humanities cannot coincide with its praxis. It can, as a human and social activity, very probably learn a lot from older theories - the relationship of actions to power structures comes to mind - but first of all it must be founded in a very close look at the activities of digital humanists, especially if it wants to come close to the original meaning of theory. In the following paper we will discuss data modeling as an abstraction of many practices in the digital humanities and we hope this will be understood as a contribution to a theory of digital humanities.

Defining Data Modeling

The question of how to define data modeling is of course of central concern, not only as a question of clarification of terms, but also as a larger question of how to situate data modeling in the appropriate context. The situation is troubled at the outset by the fact that the term "data modeling" in computer science is most typically used in a fairly restrictive sense for the modeling of relational databases, while the digital humanities has a more general understanding of the term: data modeling is the modeling of some segment of the world in such a way to make some aspects computable, referring to creating database schemas, SGML DTDs, XML schemas, ontologies etc. Thus while one can learn a great deal from the extensive discussion of data modeling in computer science, the task at hand is to define and study the more general concept.

In computer science, data modeling is "a collection of conceptual tools for describing data, data relationships, data semantics, and consistency constraints"[3] Another definition puts stress on the logical side: "A *data model* [...] is an abstract, self-contained, logical definition of the data structures, data operators, and so forth, that together make up the abstract machine with which users interact."[4] Data modeling is thus understood to consist of a set of steps. The first is *conceptual data modeling* — identification and description of the entities and their relationship in the 'universe of discourse' — so the established term for that part of the world a modeler is modeling, and notation of the findings, for example in an entity-relationship-diagram. The second is *logical data modeling*: defining the tables of a database according the underlying relational model. The third is *physical data modeling*: optimization of the database for performance, in an actual implementation. There seems to be a consensus that the third level is somewhat at the periphery of data modeling. Ideally both the conceptual and the logical model should be designed without any reference to the implementation. Thus the implementation can be

---

http://journalofdigitalhumanities.org/1-1/who-you-calling-untheoretical-by-jean-bauer/

[3] Abraham Silberschatz / Henry F Korth / S. Sudarshan (ed.): Database System Concepts. New York: McGraw-Hill 1996 (3rd edition), p. 7.

[4] C.J. Date: SQL and Relational Theory. Sebastopol: O'Reilly 2012 (2nd edition), p. 12.

optimized or even replaced at a later time. Even if the distinction between the logical and the conceptual level is a result of specific database modeling techniques, it captures an important general aspect of data modeling. The logical model provides a structure for the data which allows the user to use a set of algorithms to answer questions of interest in relation to the data. This computability is usually achieved by using a mathematical model: relation, in the case of databases, or trees/graphs in the case of XML. In these cases the logical model is a powerful formal abstraction, but it fails to represent most of the semantic information. The conceptual model addresses this lack: it captures semantic information and offers an integral and embedded view of the data, organizing the information in such a way that the logical model can either be derived automatically or is at least very easy to derive.

From the existing work in computer science and philosophy, we thus have a set of definitions and a conceptual foundation for discussing data modeling. However, when relocating/using these concepts to discuss data modeling in the digital humanities, some points of friction emerge and also some issues that require more detailed scrutiny. In the workshop on which this white paper reports, the definitions of data modeling that emerged from within the digital humanities frame of reference were much less clear-cut. The participants were deliberately chosen to represent fairly diverse fields, including linguistics, scholarly editing, history, visual arts, game studies, classics, philology, literary studies, and geography. For many participants, given that "data" could be considered as whatever formal information one might have in one's research landscape, and "modeling" could be considered as any act of formal structuring, "data modeling" thus could be defined in a loose sense as the set of formal representational activities through which information is constituted in digital form. In this sense, it was suggested that "data modeling" and "information modeling" might be at some level interchangeable terms.

However, the discussion also involved numerous attempts to add precision to this definition—to distinguish data modeling from other representational activities—and also to explore some related questions about how a concept like "data modeling" might function distinctively within the field of digital humanities. Several points in particular are worth mentioning here.

1. The effect of user requirements on the data model. It is a common feature of literature on data modeling that in order to create and evaluate a model one has to have a clear understanding of the user requirements for the data model. In the workshop discussions we noted an interesting duality in this respect. On the one hand, data models serve as an interchange format for some types of users and user communities where data is typically being created and modeled with someone else's needs in mind (archivists, libraries, others whom we might characterize as "altruistic modelers"). On the other hand, data models also exist whose function is to express specific research ideas in cases where data is being created to support the creator's own research needs (particularly for individual scholars and projects, whom we might characterize as "egoistic modelers"). Altruistic modelers also make assumptions what features of the digital objects are of interest for most users and in most use cases, while egoistic modelers can and will concentrate on their own needs. Thus, we have in practice very different ways of modeling, one trying to include very different views on digital objects and aiming to establish standards (and this involves very specific processes to decide on these user needs and to connect these new models with existing traditions of modeling, for example in library science), while the other is interested mainly to express as exactly as possible the theoretical assumptions and research interests of one or more scholars. Often the data model for research purposes is evolving during the research process and many functions of schemas are not that important in such a context.[5]

[5] One could capture this as the difference between representational markup and heuristic markup (as

2. Discipline and terminology. Another area of additional complexity arises from the disciplinary breadth of digital humanities, and the consequent flexibility of terminology used to describe the object being modeled, with consequences for our understanding of what information is to be modeled, and what the model is for. Terms such as "text", "document", "model", "transcription" , and even "data" are used in humanities discourse in flexible ways that reflect both different conceptions of these terms (arising from different disciplinary contexts) and also in some cases long-standing debates within specific disciplines: for instance, the relationship between "text" and "document" within traditions of textual scholarship. To speak of "modeling a text" requires us to specify whether we consider a text to be a linguistic object, a material object, a conceptual object (akin to the FRBR "work"), or even in some cases a string considered apart from its possible linguistic properties. In developing concepts of data modeling in the humanities, for some purposes we thus need to move from a colloquial and implicit understanding of such terms to a more precisely and explicitly elaborated set of terms and definitions for use in contexts where definitional distinctions matter.

3. Data modeling and "just plain modeling." As the workshop discussion revealed, the term "data modeling" in the humanities stands in an unclear relationship to the concept of "modeling" more generally. As Willard McCarty's exploration of the term in "Knowing … : Modeling in Literary Studies" suggests, it is possible to understand modeling as an epistemological activity: "use of a likeness to gain knowledge of its original," and this sense of the term was common in the workshop discussion.

Is "modeling" on its own something distinct, or is all modeling is ultimately data modeling? Two positions emerged. One held that there is no difference between modeling and data modeling and that any kind of modeling in the humanities is data modeling. Wendell Piez was a proponent of this position and his keynote presentation explored it in some detail. In adopting this position, we need to consider whether there is a real world accessible to us without modeling; perhaps all we do is modeling (and if so we need to think about the consequences of thinking/working this way). On the other hand, others felt that data modeling is a specialized kind of modeling, aimed specifically towards computational processes and therefore subject to more formalized constraints.

In this sense, data modeling in digital humanities can be seen as a point of connection between two intellectual domains that are traditionally considered separate: a "higher" domain in which we acknowledge an interpretative relationship with an artifact, and a "lower" domain in which there is process modeling, data structures, etc.

4. Data modeling and Classification. Data modeling is — in some respects — a classification task:[6] it relies on the clear definition of classes (or "entity sets" as they are termed in computer science). Classes are defined by a list of attributes shared by all members of the class. Members of the class ("entities") are defined by these attribute values. But in contrast to a philosophical ontology these definitions are not complete and are not meant to be. They contain just those features which are necessary to fulfill the user requirements the data model is designed for. The class 'person' in an address database, for example, is easily described with attributes like first name or phone number and avoids thus the pitfalls of a full-fledged personhood definition.

Chris Meister noted in the discussion).

[6] Thus it is especially appropriate that the Blackwell Companion on Digital Humanities contains two essays, one on database design with an introduction to the relational model and entity relationship diagrams by Stephen Ramsay and one generic essay on classification by Sperberg-McQueen.

Coming to this discussion about classification from cognitive science, one cannot help noticing that the undisputed model of a class (in computer science) is that of the 'Classical Theory'.[7] For some decades now there has been an intense debate about how humans create and use concepts, and the heavily disputed Prototype Theory is one of the most important proposals to overcome the shortcomings of the Classical Theory. According to Simsion 2007 these discussions have not found their way into data modeling in computer science.[8] Especially in cases where the main task is to model artefacts which are organized into classes by their users (an activity common in digital humanities), the application of a class concept following the requirements of the Classical Theory will result in a lot of unsatisfactorily borderline cases which could be avoided by following a principle of organisation which allows for instances of a class which don't show all features deemed necessary but is looked upon as a representative of the class anyway, famous examples are the sparrow in comparison to the ostrich.

With these considerations in mind, we can attempt to arrive at a definition of data modeling that responds to the special circumstances of the humanities, while also taking advantage of the rigor and precision of definitions arising from computer science. A few points must be made at the outset. **First**, when we speak in the following of "data modeling", we are referring to the models we use to shape digital surrogates and born-digital objects. While data structure is the more technical term referring to the way that the data model is represented in the internal or external memory of the of the computer, 'data model' refers to the conceptual and logical view. These views can be written down in a formal notational system like an entity relationship diagram (conceptual view) or a schema (logical view) but this diagram or schema will only make sense with an accompanying prose text defining the entities and the relations. **Second**, data models have three main functions. The first two have to do with communicating with the computer. First, the underlying logical model allows specific operations on the data; and second, data models allow us to constrain the kind of data allowed at specific points of the model thus ensuring the consistency of the data in regard to the conceptual model and in respect to the operations on the data. The third function is a kind of social praxis, not just for the computer: it allows us to communicate about the data (i.e. express our ideas about the structure of the data that we think is important within a specific mode of discourse being modeled). Both aspects contribute to the fact that the data model as a system of consistency constraints plays a crucial role in retaining the semantics of the data.

Third, data models describe (in a more or less formalized way) structures of data, so there is a difference between the data and this information structure. From computer science we can borrow the distinction between

- a **modeled instance** for example the structure of a text expressed in some markup; or an address book organized as a table
- a **data model**, for example the schema the textual markup is conformant to (like TEI), or the structure of the table

[7] "Most concepts (esp. lexical concepts) are structured mental representations that encode a set of necessary and sufficient conditions for their application, if possible, in sensory or perceptual terms." Eric Margolis / Stephen Laurence (eds.): Concepts. Core Readings. Cambridge, Mass.: MIT Press 1999, p. 10.

[8] Cf. Simsion 2007, p. 79.

- the **meta model,** for example XML, as a specific way to express information structures, or the relational model)

At least in the XML world the relationship between the modeled instance and the data model can vary considerably: the modeled instance very often instantiates just one of many very different relationships of the elements. That is to say it belongs to the class described by the data model while the class may contain many instances which can differ considerably from the structure described if one would construe a schema just based on the modeled instance.

And **fourth,** we must consider ontologies. The relation between the concepts 'data model' and 'ontology' is uneasy at best, not least because some use 'data modeling' as referring specifically to a relational database technique while some, but only very few, use it as a broader term. Looking at the following definition of ontology by Tom Gruber, who was one of the first to use the term "ontology" in the new computer science meaning (as opposed to the established meaning from philosophy, i.e. the science of being), one can easily see that it is almost identical with conceptual data modeling as described above:

> In the context of computer and information sciences, an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application. In the context of database systems, ontology can be viewed as a level of abstraction of data models, analogous to hierarchical and relational models, but intended for modeling knowledge about individuals, their attributes, and their relationships to other individuals. Ontologies are typically specified in languages that allow abstraction away from data structures and implementation strategies; in practice, the languages of ontologies are closer in expressive power to first-order logic than languages used to model databases. For this reason, ontologies are said to be at the "semantic" level, whereas database schema are models of data at the "logical" or "physical" level.[9]

However, there are two key differences between an ontology and a conceptual model expressed in a ER-diagram. First, as Gruber notes, "Ontologies are often equated with taxonomic hierarchies of classes, class definitions, and the subsumption relation" . And second, ontologies are designed "for the purpose of enabling knowledge sharing and reuse"[10]—in other words, they are explicitly intended as a form of communication and interchange, rather than as an internal part of a local system. (Perhaps ontologies are essentially altruistic data modeling in its purest sense.) As we are using the term 'data modeling' to refer to a general notion which encompasses all forms of modeling of data, the concept of ontology becomes a subclass referring to those models which have a very specific user requirement: represent a conceptualization of a domain that is commonly agreed to by most parties and that is relatively task

---

[9] Tom Gruber: Ontology. In the *Encyclopedia of Database Systems*, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag, 2009 (online)

[10] Both quotations also by Tom Gruber but from his famous first definition of 'ontology' and the following clarifications 1992. (online) The second point is repeatedly emphasized in CS literature, see for example Peter Spyns / Robert Meersman / Mustafa Jarrar: Data Modelling versus Ontology Engineering. 2002. (online)

independent.[11] Based on these two constraints we could say that ontologies belong to the more general class of data models, but are restricted to the conceptual level: the entities are often organised in taxonomic hierarchies and the main user requirement is enabling knowledge sharing and reuse. On the other hand almost all data models do include an ontology - at least in a weaker sense - as long as there is a conceptual level and its classes can be organised in a taxonomic hierarchy and the representational primitives are not meant to be a part of a private language.[12] And for some, it is possible to see on each level a specific ontology:

> Since an ontology is a model of a domain describing objects that inhabit it, all three types of data models can be thought of as ontologies. They range from the most expressive one that describes business concepts and processes (the conceptual model) to less expressive and progressively moving from describing business semantics to describing physical structures of the data as it is stored in the databases (the logical and physical data model).[13]

5. Complicating the idea of logical and conceptual modeling

And a last clarification: The phrase "data modeling" is used to describe two activities, closely related:

- the process of creating a data model.
- the process of applying a data model to data in order to create a modeled instance.

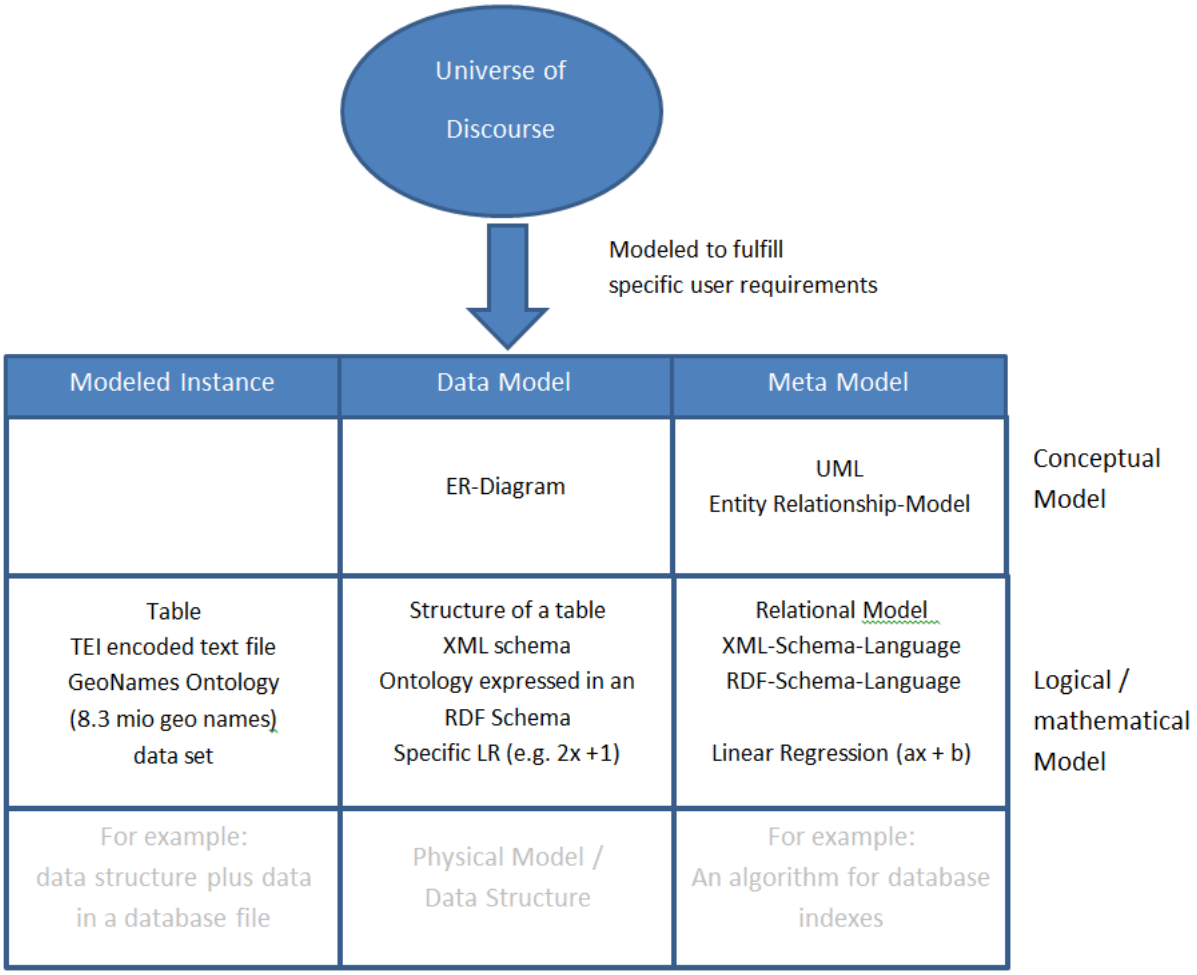
In the following discussion, we are usually talking about "data modeling" in the first sense, of "creating a data model", and we flag explicitly any cases where we use it in the second sense.

The following diagram integrates many of the distinctions discussed above:

[11] Cf. Tharam Dillon et al.: Differentiating Conceptual Modelling From Data Modelling, Knowledge Modeling and Ontology Modeling and a Notation for Ontology Modeling. 2008. (online)

[12] If you use data modeling in the relational database meaning then another difference is important: databases usually make an closed world assumption, that is all rows of a database can be seen as predicates and only these are true for the universe of discourse represented while all others are false. Ontologies on the other hand usually make an open world assumption, that is knowledge not included is unknown and not false.

[13] Irene Polikoff: Ontologies and Data Models - are they the same? Blogpost 2011 (online)



In the XML world the logical model can be expressed with an XML schema, but there is no common conceptual model for an XML schema. Maler and Andaloussi propose a tree diagram which has some of the functionality of an ER-diagram.[14] There have also been some attempts to extend existing technologies like UML or ORM (Object Role Modeling) but most have difficulties to cover all aspects of XML.[15] In the TEI context the ODD format[16] has been developed which contains the XML schema (fragments) and the prose text explaining the tags and attributes in their context. The format lacks a graphical representation

[14] See Eva Maler and Jeanne El Andaloussi: Developing SGML DTDs: From Text to Model to Markup. Prentice Hall 1995, Appenix B.

[15] A comprehensive overview can be found in Haitao Chen / Husheng Liao: A Survey to Conceptual Modeling for XML. 2010. (online)

[16] "One Document Does it all," a literate-programming format that represents the TEI schema in a conceptual, documented form from which an actual schema (in one of a number of different schema languages) can be generated.

and the abstraction from the logical model which usually can be found in the conceptual model but preserves some of the functionality of a conceptual model. As mentioned before there is a marked difference between the logical and the conceptual model: the first is an abstraction of the latter, contains less information because this part of the conceptual model can only be expressed in natural language. For the XML world there have been attempts to overcome this limitation by describing a 'formal tag-set description' which would allow inferences based on the semantics of the markup, but the problems inherent in XML and the common usage of XML are formidable (Sperberg-McQueen/Huitfeldt 2011).

Data Modeling in the Humanities

Identifying the relevant entities in the world of discourse, discerning their relevant features, describing their connections, all this is at the core of data modeling in general - relevant always in regard to the user requirements. In the Humanities very often these entities have been discussed at length as concepts either on the object-level or the meta-level. On the object-level humanists reconstruct the use of a term at a specific location and during a specific time; the history of ideas for example is a prominent field of research interested in this kind of reconstruction. On the meta-level humanists define and use terms to classify objects, for example when they attribute literary texts to an epoch (such as Realism or Baroque) or a genre (such as the Bildungsroman). In both cases the construction or reconstruction of these concepts is embedded in a longer history of discussing them, trying to capture their salient features and to find definitions. These definitions are more or less strict and thus can be formalized more easily or not at all but they almost always contain much more information about the entities than the formalization can capture. And because of the historical nature of most of the entities and because of its tendency to self-reflexion researchers in Digital Humanities view the intellectual work on this layer as an important work on its own.[17] On the other hand, even a set of very strict definitions may be not formalized enough to be the basis for a *data* model and thus demand the extra work to formalize them. It is this embeddedness which makes brute force approaches which are often the first step in computer science so unsatisfactory for cultural objects.

Data modeling can be seen as an activity involved in many different activities in digital humanities: for example, creating databases to capture informational details of cultural objects, creating digital editions by using text markup to represent the structure of text and witness information, creating software for research purposes to work on specific data sets. All these activities are similar in that the researcher has to decide what features of the object are important enough to invest time to make them explicit, and how to describe and relate these features to each other into some general structure. The results of data modeling

[17] This can be seen in projects as different as Elaine Svenonius's description of the theoretical foundations of cataloging and the prose in the TEI guidelines; see Svenonius, Elaine. *The intellectual foundations of information organization*. Cambridge, MA: MIT Press 2000; TEI Consortium, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

can be found in manuals, schemas, database designs, software designs, stylesheets and many other places; often they are not very well documented and not part of the explicit description of a project. People working in the digital humanities come across data modeling in these different ways and there is no general understanding that these activities belong to the same class and should be understood to have many commonalities our discipline has to identify, to collect, to accommodate to new ways of thinking and to teach.

In particular, we are interested in thinking about how the term "data modeling" would operate within a humanities context, and whether there is anything specific about modeling in such a context: whether because of special properties of humanities research objects, humanities methods and practices, or humanistic approaches to working with data. In the workshop discussions, several important features of humanities research objects were noted:

- they are usually artifacts rather than natural objects; human agency has played a role in their making or selection;
- they are created with a purpose and an audience;
- their history and provenance is part of their identity.

These three features all contribute to our understanding of such objects, but they also contribute to the field of information that we expect to model about these objects in order to create adequate representations of them for research purposes.

In addition, there seems to be a marked difference to the common understanding of data modeling in the Digital Humanities. In computer science, both theorists working in the academy and those working in industry doing practical data modeling, most regard data modeling as a description of a real and objective world (which includes the possibility of assessing the correctness of data models) while only a minority views it as a design process.[18] However, in digital humanities there seems to be a general understanding that a data model, like all models, is an interpretation of an object, either in real life or in the digital realm. Michael Sperberg-McQueen in his closing keynote to the workshop stated this position clearly: "modeling is a way to make explicit our assumptions about the nature of a text/artefact." Furthermore, most digital humanities researchers assume that data modeling is primarily a constructive and creative process and that the *functions* of the digital surrogate determine what aspects have to be modeled.

There is both an ontological and an epistemological question here. The ontological question concerns the presence of a common objective (or at least intersubjective) reality. Such a reality is difficult to establish even for material or natural objects, and it is even more difficult if we include all of the entities that play a

[18] See Graeme Simsion: *Data Modeling: Theory and Practice*. Bradley Beach: Technics Publications 2007, who bases this statement on the analysis of computer science literature on data modeling and a broad set of interviews and questionnaires. Simsion's book (Simsion 2007) is centered around this distinction and states: "definitions of data modeling commonly characterize it as a description, but there are dissenting views and common metaphors which align better with the design characterization." Simsion 2007, p. 32.

larger role in data modeling, such as audience, purpose, date, etc., all of which are social constructions. The epistemological issue concerns whether it is possible to construct multiple, equally valid views of a socially constructed reality. Most modern-day researchers would probably agree that it is, because we have learned in the last hundred years how each view is entrenched in a complex set of preconditions everyone has acquired. With these two aspects we have a sound basis for a generally shared intuition that on the one hand data modeling is a way to make a specific view of the world explicit, and therefore there can be many equally valid views—but on the other hand it is also possible to speak of more or less effective or useful data models, and perhaps also of useless or wrong data models as well. Some data models cover a socially constructed class in ways which seem to most observers within a research community to represent more features in a more elegant or economic way than others.

Based on these observations we offer our own definition, applicable to the digital humanities: Data modeling refers to the activity to design a model of some real (or fictional) world segment to fulfill a specific set of user requirements using one or more of the meta models available in order to make some aspects of the data computable and to enable consistency constraints. Two typical forms of user requirements can be found: (i) one which aims to define a data model for a domain in order to support data exchange and long-term use. (ii) Another typical form of user requirement is mostly interested to model specific research assumptions and is often only used for a short time in a specific research context. In general a data model consists of (i) a conceptual part which defines the data semantics, the relevant entities, their features and their relations, and (ii) a logical part which expresses a subset of the conceptual model in such a way that it is an abstract, self-contained, logical definition of the data, data operators, and so forth that together make up an abstract machine which makes the data computable. For technical reason these two parts are usually expressed as two different models in data models of today, but there is no inherent reasons that this will also be the case in the future.

2. Anatomy of Data Modeling

Creating Data Models

The process of creating a data model is well known in the digital humanities as a practical activity, especially in the area of schema development.[19] In contrast with computer science, where 'data modeling' refers almost exclusively to database design and to object-oriented modeling in software systems,[20] [21] the more

[19] See for example Maler and Andaloussi.

[20] The discussion on data modeling is rather advanced in computer science; one interesting area is the attempt to describe underlying patterns of data modeling, parallel to describing patterns in programming; cf. for example Len Silverston, Paul Agnew: *The Data Model Resource Book* Vol. 3. Universal Patterns for Data Modeling. John Wiley 2009.

[21] There is a huge amount of literature on data modeling in the computer science meaning created by and adressing the needs of practitioners.

practically oriented discussion in digital humanities is centered around recurring problems. Some of these are fundamental problems, such as the inherent conflict between standardization and expressiveness, or the challenge of balancing functionally driven and theoretically driven modeling imperatives, or the strategic question of how to determine when a model has been sufficiently tested against the target data. Some are problems created by the evolution of the digital ecosystems the objects live in; for example, the existence of linguistic corpora each using a slightly different approach to modeling linguistic features like part of speech created the need for new models allowing retrieval of the same kind of object independent of its project-specific label. Others are problems arising in connection with specific tools or activities, such as the fact that the use of XML entails the development of workarounds for the problem of overlapping hierarchies, or the question of how to provide for durable annotation of dynamically changing data. In digital humanities, we might say that data modeling is not primarily understood as an activity that takes place prior to building systems, but rather is understood as being deeply involved with a variety of processes and tools that are implicated throughout the life cycle of data. The recent growth in emphasis on data curation has given even greater visibility than before to the idea that data models are created and reworked in the process of tool design, data capture, documentation, interface design, publication, archiving, and reuse.

To advance our understanding of data modeling in digital humanities, then, we should take a closer look at these practical contexts and at the ways in which they reveal opportunities for theorizing these modeling practices more rigorously and explicitly. In the following sections we consider the context of data modeling, the practice of data modeling, forms of notation and documentation and the role they play in shaping the semantics of data models, and finally tools for data modeling.

The Context of Data Modeling

Data modeling in the humanities usually means to work on a type of entity or object that has a history, often a very complex one. The class of documents we term "letters," for example, has a long and manifold history which over time has changed in many aspects, both in the nature of the material object and in the structure of the text. Furthermore, while we can consider a "letter" as a text (in the linguistic sense) and model it as such, there are aspects of layout, use of the writing space, and the arrangement of words on the page that are so important to our understanding of the letter's meaning that we may find it valuable to model it as a graphical object. Data modeling in the humanities is always happening in the context of former attempts to model either these specific classes of objects (for example letters) or the more generic class (texts, graphical objects). Therefore the history of attempts to describe these classes should be known and understood in the community, and indeed constitutes an important strand of expertise for those undertaking the modeling. And this history predates the introduction of digital tools: for example the reflection of philologists about their objects and the establishment of practices to describe them goes back into antiquity. There is thus an important historiographical aspect to modeling, considered as a record of intellectual practice; the discussion of the shortcomings of traditional or digital models (as exemplified, in

the workshop, by Allen Renear's discussion of the logical problems of FRBROO) is an important way to contribute to data modeling and digital humanities in general.

All this knowledge about these historical artefacts constitutes the context of all data modeling of these objects. It is one of the main tasks of the humanities to construct, collect and organize this knowledge, and the humanities has amassed a huge amount of knowledge about its objects, and has made systematic attempts at formalization and definition of terms. Sometimes, as in the case of library science, this has taken place in a very formal way with rather exact concepts and the use of controlled vocabularies. In other cases, as for example with the tradition of scholarly editions, this formalization has taken place in a looser sense, operating more to define concepts and practices than to define specific terms with complete precision. At the moment, very little of this knowledge is expressed formally enough to be processed computationally. This is understandable; a formalized knowledge organisation is far from easy to achieve, as we can see if we look at the discussions on FRBR and FRBROO[22] or other proposals to describe systematically even very local areas of this vast landscape under a specific perspective of usage. However, if the digital humanities is to realize the potential of this kind of historical and contextual knowledge as part of our data modeling practice, such formalizations will be increasingly necessary. And any data model in the digital humanities ignoring this knowledge risks overlooking important aspects of the object. This is especially true for what we have called "altruistic models," because in these cases older models can be regarded as descriptions of user requirements.

On the other hand there is a noticeable gap between the older, traditional, analog-world models and data models of the same objects. One reason is, as we already mentioned, the need for a stricter formalization in order to avoid the ambiguity and polysemy of natural language, which people can resolve so easily in most cases but computer still cannot. Another reason for the gap are the new requirements created by the new abilities and features of digital objects like complex searching or advanced forms of visualization. And last not least relying too much on older descriptions of cultural objects and their features "can perpetuate existing views of data"[23] instead of highlighting new features visible now in the context of the digital media.

The Practice of Data Modeling

The issue of data modeling in practice received less attention overall in the data modeling workshop. In our summarizing discussion, we noted the complexity of the human/work process of creating the data model, but this was almost invisible in the presentations. We also noted the importance of understanding the relationship between the complexity of the data and the process of developing data models; this is an area for further attention. We need to think about how to elicit this information, and also about where it lives (in many cases, it is not even visible in project documentation).

---

[22] Functional Requirements For Bibliographic Records, see http://www.ifla.org/publications/functional-requirements-for-bibliographic-records/

[23] Cf. Simsion 2007 p. 71.

On the other hand, data modeling has been a key area for IT professionals in the last decades and a lot of books have been published by computer science researchers and even more by practitioners which made this an important source of concepts and ideas to all related fields.

Thus the practical steps of data modeling have been described repeatedly, but usually not in a generic way but rather in respect to a specific metamodel, as for example Maler and Andaloussi 1996 for textual models expressed in SGML or Oppel 2009 for database design.[24] Some concepts and ideas seem to be widely shared, for example that the first step in data modeling is to write up a clear statement of the goals of your project: "The first step in designing a bibliographic system is to state its objectives."[25] Every aspect of the data model should be related to this set of objectives and this relation works in both ways: Are all objectives covered by the model and are all features motivated by one or more objective? It is a common problem that the users often don't have clear understanding of their requirements, especially in the Humanities where the socialisation of many is still mainly shaped by analogue media and the quick development of research possibilities of digital media in this area.

The following diagram shows a summary of the assumptions how data modeling is working as a process:[26]

[24] For example Andy Oppel: Data Modeling. A Beginner's Guide. McGraw-Hill/Osborne Media 2009. The example is rather arbitrarily.

[25] Elaine Svenonius, *The Intellectual Foundations of Information Organization*. Cambridge, MA: MIT Press 2000, p. 15.

[26] Taken from Simsion 2007, p. 35. Simsion claims that this diagram shows those steps mostly shared by data modeling descriptions in the computer science research literature and the literature written by practitioners.

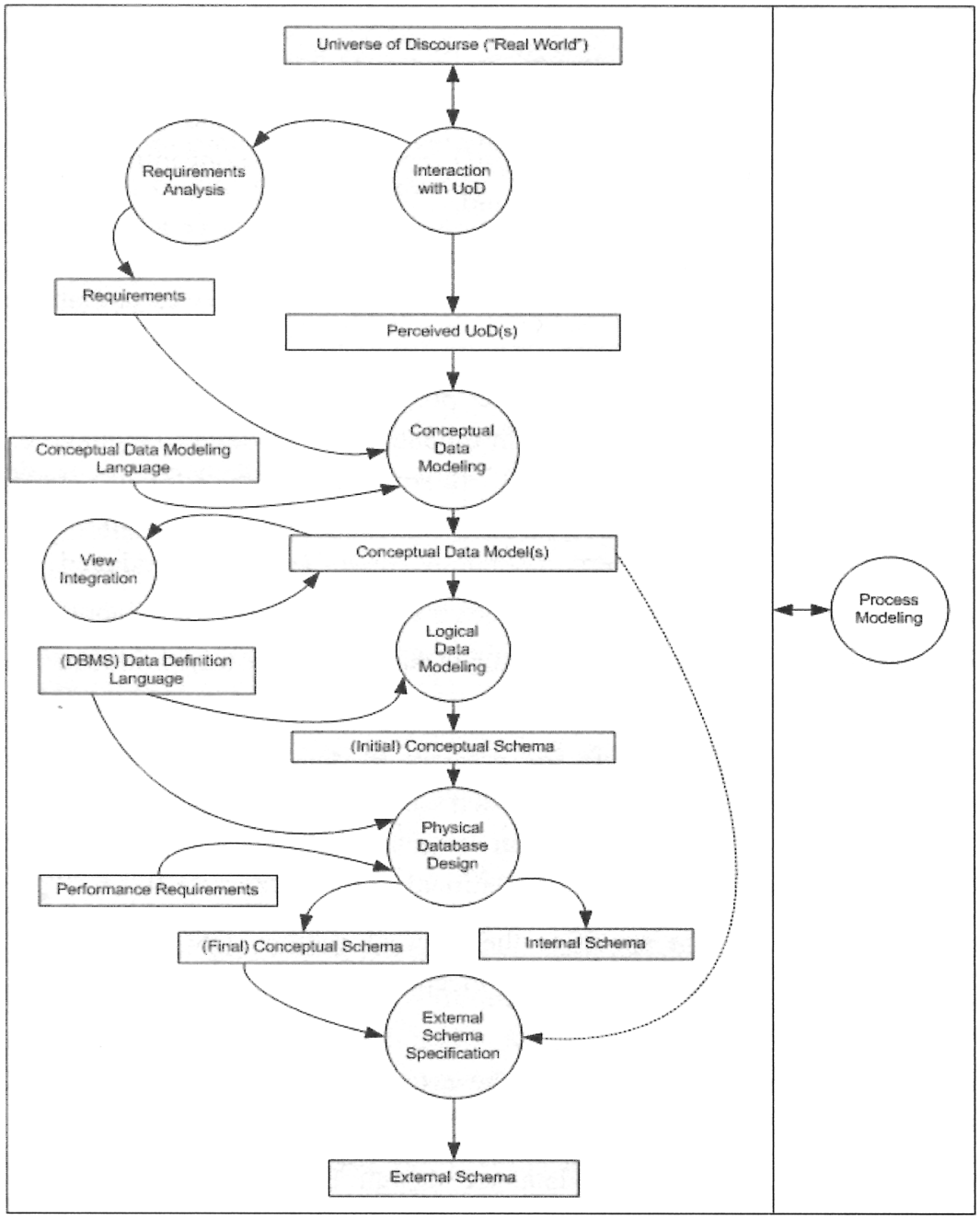


Figure 3-1: Stages in database design – a generic framework

'Universe of Discourse' refers to that part of reality which the data model tries to describe or which it constructs. The different views of this Universe of Discourse will be expressed in conceptual data models using a conceptual data modeling language and in the next step these views will be integrated into one conceptual data model. Based on a specific database data definition language this model will be expresses as a logical model which in turn will be the basis of the physical database design process resulting in the external schema. Data modeling is during all these stages informed by the process modeling. It is interesting to note that according to Simsion's overview of the data modeling literature, practitioners use

the term 'Conceptual Data Model' to refer to the scoping model, obviously not yet expressed in a data modeling language which mirrors the experience in the digital humanities, that many concepts which one wants to express in a formal data modeling language are so complex that an informal description is the first step to model them.

Though most of the experience has been gathered in designing data models for enterprises and businesses there can be no doubt that digital humanists can learn a lot from this expertise and insight, even if they have to accommodate the principles and ideas to their area of research and modeling.[27] A closer collaboration and a organized transfer of knowledge would probably minimize the need for unnecessary double research and reinventing the wheel. But a closer look into this literature also conveys a warning: for many the limits of their data modeling language and its notational system also limit what is worthwhile knowing. Humanists have a vaguer and much more embedded understanding of their objects. Even if the computer systems of today cannot handle this kind of information structure gracefully, it should be kept in mind that these are the requirements we ultimately want to satisfy.

Another important aspect coming out of the practical experience with creating and using data models and data conforming to data models is the need to distinguish clearly between declarative and procedural aspects and to exclude the latter from the data model because it has proven to much more situational, and more time- and context-dependent.

Notation and Documentation or the Semantics of data models

As already mentioned above, the formalization in data modeling only affects some aspects of the modeled concepts. Looking at the most common data models in Digital Humanities (databases and XML encoded texts) it is easy to see that there are three predominant types of information in the model:

1. Features. In XML this kind of information is usually encoded in attributes. In a logical database model it is often encoded in the headings of the columns.

2. Relations. In XML this kind of information is usually encoded in the schema, which describes where an element is allowed and which elements it can contain. The class system in an XML language like TEI adds another layer of relation semantics to this. In an ER diagram, relations are explicitly marked by arrows and their semantic is described using ordinary language. Most of this semantic is lost in the logical model which mainly expresses the relation using primary and foreign keys.

3. Data types. These contribute to the semantics by limiting the kind of data which can be input at

---

[27] A similar opinion is expressed by Ronald J. Murray in his discussion of the shortcomings of the FRBR model; cf. Murray, Ronald J.: The FRBR-Theoretic Library: The Role of Conceptual Data Modeling

[27]In Cultural Heritage Information System Design. In: Proceedings of The Fifth International

[27]Conference on Preservation of Digital Objects. iPRES 2008. p. 172-177. (Online)

specific points of the data model. Obviously this only contributes to the semantics, usually of the features, and often only a small amount of information.

Usually the largest part of the semantics is stored in the natural language headings, descriptions and definitions coming with the data model. For example it is the main goal of the ODD file format, developed by the TEI, to formalize this relationship between these parts of the data model. Another large part of the semantics can be found in the processing model, especially as it is described by the processing software, for example a set of stylesheets or a database application. (As Stephen Ramsay argued in his presentation at the workshop, "Every data model is asymptotically approaching a processing model.") Maybe or probably computer processing will be more intelligent in some future, but as long as each data model is basically an island and the context giving it its semantics is brought in by the humans, data models have to be addressed to two recipients at the same time: humans and machines. This affects its notation: it belongs to the basic principles of SGML to use tag names which can be understood by humans and the same is true for table names etc.

In the digital humanities it is an unsolved problem how to integrate the different information about the semantics of a data model stored in the different places like encoding, metadata, processing instruction, stylesheets, documentation etc., into one larger survey. This is especially true because of the widespread use of text encoding which doesn't use anything similar to an ER model. This problem is not unique to the digital humanities: both computer science and digital humanities have to work with notation and constraint, and both fields hit the wall at the same place: everyone has the same problem with semantics. However, digital humanities arguably experiences the problem more acutely because we deal in semantics above all.

Tools

The tools we use for creating and processing data are another place where data models are created and instantiated, and this fact is a revealing source of conflict. Humanists feel intuitively on the one hand that tools ought not to influence intellectual processes (we are not slaves to our tools; our ideas are independent of their material instantiation) and also, on the other hand, that the material circumstances of work (scholarly work, artistic work) do constitute a set of constraints that influence the final product. Digital humanists, and particularly those working in the tradition of text markup and open standards, feel strongly that data should be tool-independent, and that the significant information carried in a specific data format should be convertible without loss into other formats and should be portable from tool to tool. In the XML world particularly, where concepts like "archival formats," "transformation," and "single-source publication" are so deeply entrenched, tools are understood as convenient but temporary carriers whose specific requirements are accommodated by temporary, reversible transformations; adapting the data to the tool (or allowing the tool to exercise a shaping influence on the data) is generally considered illegitimate and at best an unhappy compromise. Along similar lines, we may observe the deep-seated premise in the XML world that markup should describe information constructs rather than prescribe processing (the "descriptive" vs. "procedural" debate [see Renear 2004]).

Evaluating data models

In general there seems to be a consensus that data models have to serve functions specified by the user requirements and that one important aspect for their evaluation is how well they serve these functions. This aspect poses – in theory at least – no problem and it comes down to problems of communication between a data modeler and the domain specialists who are supposed to work with the model at the end. But especially with "altruistic" data models we always face the dilemma between standardization and expressiveness or in other words: we have to deal the problem that the better a model suits one specific case the worse it will

Data models become more robust if more people do describe their requirements during their design; in these cases data models will cover more use cases and will be applicable to more situations. But making provisions for specific user needs usually increases the complexity of the model. And though a more complex model will be able to cover more of the user requirement of a specific project in that domain, most probably it won't be able to satisfy all needs. Very often the more general a model is in its coverage, the higher the probability a user will find it lacking: it won't be able to express differences in such details as seems to be necessary for her or his needs. Obviously the more specific a data model is and the better it expresses the requirements of one user or one project, the more difficult it will be to express the categories of another group of users. Interestingly most users won't refuse to work with an ill-fitting data model but will try to find workarounds for their problems by using a category manifestly meant for a different purpose, in the field of text encoding for example, this is called 'tag abuse'.

While fulfilling the user requirements seems to be the most important aspect there is also another which we would call tentatively the problem of 'truthfulness' or 'adequacy' of a model. Most participants of the workshop agreed on the role of social construction, convention, interpretation: data modeling involves making explicit an interpretation of an object. Discussion came repeatedly back to questions about level of controversiality, domains in which interpretations are widely shared, vs. less widely shared. There was a continuum with the radical position on one end ("everything is interpretation") and the less radical position on the other end ("maybe so but we basically can agree on some things like paragraphs") and interesting discussion concerning where things get problematic or controversial (post-structuralist controversy). On the other hand there seems to be an understanding that in altruistic data modeling the way an object is modeled has to conform to the social construction of this object and often a fruitful way to access this social construction is a closer look into older codifications of descriptions, for example standards of book cataloguing. In this perspective not the truth of a model is the measure it is evaluated by but just the question how well it captures a shared understanding of the some aspect of the world — independent of more ambitious theories of truth.

But maybe one of the more important insights is the distinction between what we called above altruistic and egoistic modeling. These two groups usually follow rather distinct approaches how to determine the user requirements for a data model, how to evaluate the models against these requirements and how they implement procedures to improve the data models. In the case of a researcher just working on his own

data, this procedure of improvement may consist of tweaking a schema based on test runs of an analysis, while on the other hand one may have a rather formal way like the Unicode group or the Special Interest Groups of the TEI.

A third important dimension of evaluation for data models can be described as inherent: We are talking about aspects which make a data model more robust in its use over time and more robust in the context of an application. As already mentioned above though in practice a data model will be dependent on aspects of the application which process the data, a data model has to be designed independent from the application. Sometimes it is difficult to distinguish between the requirements of the users and the requirements of the application which is built turn to satisfy the user requirements. But especially information which is needed for the management of a workflow or the data life cycle has – as we are looking here at technical metadata – to be distinguished clearly from other aspects of the data model. Thus the processing model and the data model have to be designed separately and the needs of the first shouldn't determine the latter (with the exception of technical metadata).

But independence of the application and the processing model is not the only feature of concern in making our data models more robust: as products of the digital world, data models are also always threatened by the rapid evolutions and revolutions of the field. An obvious source of trouble are changes on the level of the operating system (for example the character encoding of the system), the applications on which the data model is dependent, or the meta model through which the data model is expressed.

At first glance choosing a data models may look like a problem that is primarily technical, but it becomes soon evident that — because a data model is nowadays always part of a social practice — it is embedded in social practices and relations. If one creates a data model using the framework of some larger standard like the TEI, one is necessarily identifying with and adopting a community, a way to look at digital objects, a way to discuss and even to evaluate strategies - even if it takes time to understand all these implications and to be socialized into this way of doing digital humanities. If on the other hand one decides to create a new schema (or not to use a schema at all), one thereby chooses autonomy in preference to the established practices associated with the standard and the community. Both approaches have their advantages, and the decision is as much social and strategic as it is technical.